Overview

1. syntactic annotation and treebanks
   - lab work: TIGERSearch
2. lexical semantics
   - lab work: WordNet
3. Projects

Treebanks

A linguistically annotated corpus that includes some grammatical analysis beyond word-level syntactic annotation (part-of-speech)
- "treebank" vs. "annotated corpus"
  - the first has to be manually annotated or post-edited
- two syntactic frameworks:
  - constituent structure
  - dependency structure
Constituent structure

- American structuralism, e.g. Zelig Harris (1951)
- “Bracketing”: sentences consist of hierarchically embedded subparts → constituents
  - strings of words that belong together
  - constituency tests: substitution, movement, stand-alone test...
- Part-whole relations
  - e.g. a NP consists of a determiner, adjective and noun

\[ NP \rightarrow [DET \rightarrow [ADJ \rightarrow [N]]] \]

Dependency structure

- First comprehensive theory: Lucien Tesniere (1959)
- Sentence consists of hierarchically structured asymmetric binary relations between word forms → dependency relations (connections)
  - governor, dependent(s)
  - closely related to functional analysis
- Relations
  - e.g. determiner and adjective are subordinated to the noun
  
  \[ \text{DET} \rightarrow \text{N} \quad \text{det} \quad \text{attr} \quad \text{DET} \rightarrow \text{ADJ} \]

Dependencies in SDT
Hybrid models

Combine constituent and functional (dependency) information
- e.g. function added as additional sub-label to daughter category
  \([S \{NP-SB \ldots\}]\) in Penn Treebank II

Treebanks and linguistic theory

- **Constituent structure, e.g.:**
  - Penn Treebank I (AE)
- **Dependency structure, e.g.:**
  - Prague Dependency Treebank / analytical level (Czech)
- **Constituent / Dependency Hybrid approaches, e.g.:**
  - Penn Treebank II, SUSANNE (AE)
  - NEGRA/TIGER, TüBa (German)
- **Theory specific annotation, e.g.:**
  - Prague Dependency Treebank / tectogrammatical level - Functional Generative Grammar
  - CCG-bank - Combinatory Categorial Grammar

Penn Treebank

- English treebank built at the University of Pennsylvania, distributed by LDC: [http://www.ldc.upenn.edu/](http://www.ldc.upenn.edu/)
- Phase I (1989 - 1992)
  - skeletal parse
  - 2.6 million words PoS tagged from Wall Street Journal, also other components, e.g. Brown Corpus
- Phase II (1993-1995)
  - enriching part of the material with grammatical functions and semantic relations
  - null-elements, coreference
- Phase III (1996-2000)
  - additional material: corpus of telephone conversations annotated for disfluencies
Penn Treebank: PoS annotation

- uses modified BROWN tagset
- allows multiple tags on word when annotator is unsure (avoid arbitrary decisions)
- 36 PoS tags, 12 other tags (punctuation, currency symbols)

```plaintext
1. CC Coordinating conj. 19.TO to
2. CD Cardinal number 20.MN conjunction
3. DT Determiner 21.TN that, base form
4. EX Existential there 22.VBD Verb, past tense
5. FW Foreign word 23.VBG Verb, present participle
6. IN Preposition/subordin.
7. LS LS Linking v.
8. MD Modal 25.VBD V, base form, present
9. NN Noun, plural. 26.VBG V, base form, present
10. NNP Proper noun, singular 27.RB RB, base form
11. NNPS Proper noun, plural 28.JJ JJ, base form
12. NNS Noun, plural 29.RB RB, base form
13. NNP Noun phrase 30.RB RB, base form
14. PRP Personal pronoun 31.RB RB, base form
15. PRP$ Possessive pronoun 32.RB RB, base form
16. IN Preposition/subordin.
17. IN Prepositional phrase 33.RB RB, base form
```

Penn Treebank: syntactic annotation

1. ADJP Adjective phrase
2. ADVP Adverb phrase
3. NP Noun phrase
4. PP Prepositional phrase
5. S Simple declarative clause
6. SBAR Clause introduced by subordinating conjunction or 'that'
7. SBSAQ Direct question introduced by vb-word or wh-phrase
8. SINV Declarative sentence with subject-aux inversion
9. SQ Subordinate of SBSAQ excluding vb-word or wh-phrase
10. VP Verb phrase
11. WMOV Movew from phrase
12. WRB Wh-word phrase
13. WNP Wh-noun phrase
14. WP Word phrase
15. X Constituent of unknown or uncertain category

Penn Treebank: Skeletal parsing

```plaintext
(S
  (NP Martin Marietta Corp.)
  was
  (VP given
    (VP a
      $ 29.9
      million Air Force contract
    (PP for
      (NP low-altitude navigation and
       targeting equipment))))
```

Penn Treebank: Skeletal parsing

```plaintext
(S
  (NP Martin Marietta Corp.)
  was
  (VP given
    (VP a
      $ 29.9
      million Air Force contract
    (PP for
      (NP low-altitude navigation and
       targeting equipment))))
```
Penn Treebank: Functional tagset

- Text categories
  - HNL headlines and datelines
  - LST list markers
  - TTL titles

- Grammatical functions
  - NOM non NPs that function as NPs
  - ADV clausal and NP adverbials
  - SBJ surface subject

- Semantic roles
  - DIR direction and trajectory
  - LOC location
  - MCR manner

- Pseudo-attachment
  - "EXP" expletive
  - "RNR" right node raising

TIGER Treebank

"Linguistic Interpretation of a GERman Corpus"

- 50,000 sentences
- Follow-up of NEGRA corpus (20,000 sentences)
- German newspaper texts (Frankfurter Rundshau)
- Free licence
- Hybrid annotation
- Crossing branches for discontinuous constituents

TIGER treebank example: discontinuous constituents

Ein Mann kommt, der lächelt

A man comes, who laughs
Creating treebanks

- Manual annotation
  - TrEd, CLaRK, Word freak
- Automatic annotation with human post-editing
  - Collins’ Parser, Stanford Parser,…
- very labour intensive!

Exploiting treebanks: Parser training

```
S
  VP
    V
      drew
    CT
      the
do
    CT
      draft
draft
    CT
      her

S
  VP
    V
      she
    CT
      had
draft
    CT
      an
agreement
    CT
      with
draft
    CT
      her

S
  VP
    V
      she
    CT
      had
agreement
    CT
      with
draft
    CT
      agency
```
Exploiting treebanks: Parser training

Exploiting treebanks: Parser training

Exploiting treebanks: Charniak 1996

- inducing a treebank-based PCFG
- preliminary version of Penn Treebank
- training corpus: ~30,000 words
- test corpus: ~30,000 words

<table>
<thead>
<tr>
<th>Sentence Length</th>
<th>Average Length</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-12</td>
<td>8.7</td>
<td>88.6</td>
<td>91.7</td>
</tr>
<tr>
<td>3-15</td>
<td>13.4</td>
<td>85.5</td>
<td>97.7</td>
</tr>
<tr>
<td>3-20</td>
<td>13.6</td>
<td>83.5</td>
<td>88.2</td>
</tr>
<tr>
<td>2-25</td>
<td>16.3</td>
<td>52.0</td>
<td>64.0</td>
</tr>
<tr>
<td>3-30</td>
<td>18.7</td>
<td>80.6</td>
<td>52.0</td>
</tr>
<tr>
<td>2-40</td>
<td>21.9</td>
<td>78.0</td>
<td>50.4</td>
</tr>
<tr>
<td>CoNLL-X shared task on multilingual dependency parsing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 2006, <a href="http://nextens.uvt.nl/~conll/">http://nextens.uvt.nl/~conll/</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- open task: common format of treebanks, all systems must compete on all languages</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 13 treebanks: Arabic, Chinese, Czech, Danish, Dutch, German, Japanese, Portuguese, Slovene, Spanish, Swedish, Turkish, Bulgarian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 20 systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Best average labelled attachment score ~80%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>